Introducción al Análisis Estadístico

Dr. Delfino Vargas Chanes

Facultad de Economía
Universidad Nacional Autónoma de México

15 de febrero de 2024

Índice

Introducción

- La estadística en la actualidad se asocia con la colección de datos, levantamiento de encuestas, conteo de elementos de una población. También suele asociarse con aspectos de análisis de datos, procedimientos de muestreo.
- Una acepción más exacta el término estadística es la que se relaciona con la aleatoriedad, eventos que no pueden repetirse sin variación y esto nos conlleva a que la aleatoriedad es uno de los elementos fundamentales de la estadística.
- Los métodos estadísticos tienen relevancia cuando los eventos que muestran cierta regularidad en su realización pero que no se sabe con certeza del resultado específico que se obtiene en algún caso de estudio, a esto se refiere con el término *aleatoriedad*.

Estadística e Incertidumbre

- Cuando se colecciona información en las Bases de Datos usualmente se tiene un fin. Por ejemplo, si se busca conocer los hábitos de compra de los consumidores de una cadena de supermercados la información de ticket de compra será de mucha utilidad.
- Sin embargo, la información es finita, quizás solo se tiene información de los clientes con tarjeta de lealtad, o solo se dispone de una muestra de clientes de una cadena de hoteles que contiene datos de una encuesta de satisfacción.
- Este escenario plantea que se tiene cierta incertidumbre al obtener una inferencia al resto de la población.

¿Qué es la Estadística?

- La estadística es la ciencia que se ocupa de reunir, organizar, presentar, analizar e interpretar datos para ayudar a tomar mejores decisiones.
- Importancia de la estadística:
  - La información se encuentra disponible casi por cualquier medio, pero el problema fundamental es procesarla para entender el mundo que nos rodea.
  - Las técnicas estadísticas se usan para la toma de decisiones. La información que es útil se puede procesar para entender mejor la toma de decisiones.
  - El conocimiento de los métodos estadísticos ayudan a entender la toma de decisiones y cómo afecta el mercado.
  - Sin embargo, la información se produce con un fin específico, **hay que evitar los estudios GIGO** (garbage in garbage out).

Definición de Estadística

Por tanto, la Estadística se puede definir como:

> *"el cuerpo de conceptos y métodos que conciernen a un área específica de la investigación para obtener descripciones y conclusiones de situaciones donde la aleatoriedad esté presente."*
>
> — (Bhattacharyya y Johnson 1977)

La Estadística Descriptiva

- La parte más antigua de la estadística está integrada por un conjunto de técnicas para la organización, presentación gráfica y cálculo de cantidades **representativas** de un grupo de datos. Esta parte se llama estadística descriptiva.
- Por ejemplo, si se estudian las adicciones en México, podemos partir de un conjunto de datos que indique ciertas tendencias de la prevalencia de uso de alcohol entre adolescentes y jóvenes.
- Pero partimos de un subconjunto de datos que representan a la población. Se dispone de un conjunto de datos que suponemos representan a la población de estudio.
- Usamos el método inductivo para obtener conclusiones de la población objetivo a partir de observaciones tomada de una muestra.

Objetivos de la Estadística

- Planear la obtención de datos con un fin especifico para poder llegar a conclusiones confiables.
- Determinar si la información existente es adecuada o si se requiere de información adicional.
- Una buena planeación del estudio evita un doble esfuerzo.
- Analizar la información disponible.
- Sacar conclusiones y hacer inferencias determinando el riesgo de una conclusión incorrecta.
- Resumir la información de una manera útil e informativa.

¿Qué es la Estadística?

- El término estadística en la actualidad se asocia con la colección de datos levantamiento de encuestas, conteo de elementos de una población.
- También suele asociarse con aspectos de análisis de datos, procedimientos de muestreo. Una acepción más exacta el término estadística es la que se relaciona con la aleatoriedad, eventos que no pueden repetirse sin variación.
- Por tanto la Estadística se puede definir como *"el cuerpo de conceptos y métodos que conciernen a un área específica de la investigación para obtener descripciones y conclusiones de situaciones donde la aleatoriedad esté presente"* (Bhattacharyya y Johnson 1977).
- La palabra estadística se deriva del latín *status* que se relaciona con aspectos del *estado*. Es decir, con aspectos de colección de datos demográficos, económicos y políticos fundamentalmente.

La Estadística y la Investigación Científica

- Buena parte de la investigación científica ha surgido de corrientes como el empirismo, la fenomenología, el estructuralismo que buscan dar respuesta a la búsqueda de la verdad. La discusión a veces centra en el paradigma de la objetividad y la subjetividad y suele reclamarse que las investigaciones son científicas si son objetivas.
- El debate se ubica entre si lo **subjetivo** es inválido y lo **objetivo** es válido en el contexto de la investigación social, pero para ello debo recurrir primeramente a debates que ya han tenido lugar en el pasado y ubico este debate en el contexto de la investigación social, más que en el lado filosófico.

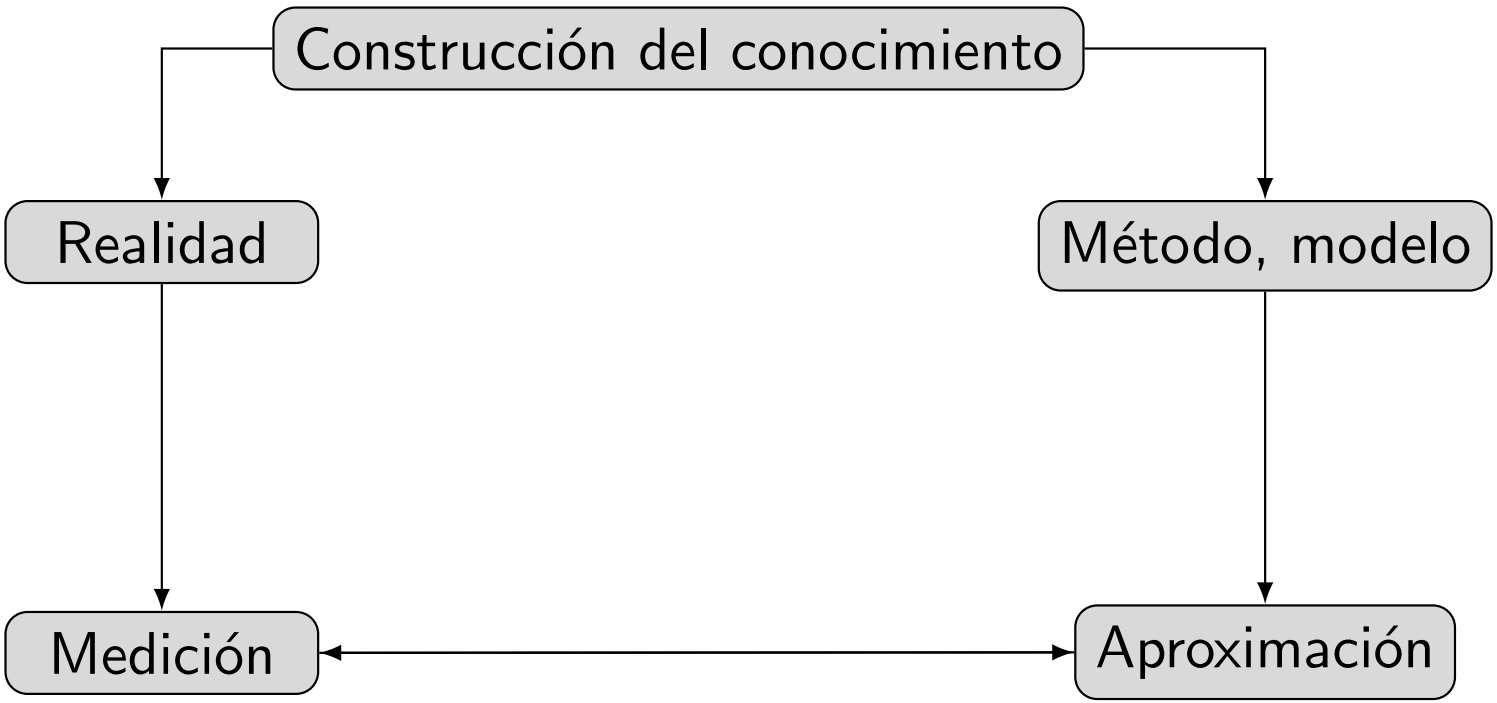
El paradigma de la investigación

- A mediados del siglo XIX se argumentaba que la investigación en **Ciencias Naturales era el paradigma** que reclamaba ser la única forma de hacer ciencia y sobre esta base la tradición positivista solo se favorece a la objetivación como la única forma de hacer ciencia.
- Se dice que, bajo este paradigma, investigadores de otras disciplinas "*duras*" postulan que **la investigación solo debe basarse en hechos objetivos**.
- La **objetivación** es el proceso que reclama que el conocimiento científico es tangible y repetible, los objetos están ahí definidos y basta con tocarlos para probar que existen.
- La **subjetivación** es el proceso que requiere de la interpretación de los hechos, los científicos observan los objetos y extraen la información pertinente para definir un fenómeno y hacer ciencia.

El método

- El método es una forma de extraer el conocimiento y la construcción del conocimiento no se debe distraer por el método utilizado, ya que solo es un medio, entonces **la parte sustancial es la construcción del conocimiento**.
- La tesis central que se plantea es que l**a realidad en sí no se conoce ni está determinada pero sí se puede organizar.** La realidad se construye a partir de la interacción entre el investigador y el objeto de estudio.
- **Nos aproximamos a la realidad** desde varias trincheras, desde las Ciencias Naturales se puede dar una aproximación, desde las Ciencias Económicas y Humanísticas se da otra aproximación ¿Cuál de todas estas es la verdadera?
- **¡Pero la realidad no es única pero sí se puede organizar!**

La Realidad y el Método



- **La realidad no es única, se construye** y lo que hacemos con los métodos y los modelos (estadísticos-econométricos) es ofrecer una aproximación desde una trinchera específica, pero no única.

¿Cómo nos aproximarnos a una realidad?

- Podemos construir una realidad mediante un enfoque subjetivo y si este enfoque es similar a otros y forman un consenso se vuelven objetivas. Pero son objetivas para el conjunto de visiones que lo plantean así. **Las verdades no son universales.**
- Ejemplo: queremos medir la cohesión social de una población objetivo, para ello se utiliza un instrumento que busca medir ese constructo que incluyan todas las posibles dimensiones de la cohesión social.
- La medición puede tomarse como subjetiva preguntando a las unidades de estudio si este proceso se mide con un error mínimo pasamos por un proceso de subjetivación al de objetivación al encontrar un consenso de las mediciones a lo largo de diferentes condiciones de variación de la cohesión social.

La realidad no es directamente observable

- La realidad es muy compleja y depende el punto de vista de donde se vea, pero la realidad es organizarle a través de métodos específicos diseñados para mostrar una visión más entendible.
- Los modelos nos ayudan a organizar la realidad y en particular los modelos cuantitativos. Ningún modelo es *"el bueno"* pero si es necesario entender los supuestos del modelo para tener una mejor certeza de que el modelo aproxima lo suficiente y para ello debemos comprender los supuestos.

Tipos de Variables

Variables cuantitativas

- Son aquellas que se pueden medir (estatura, horas frente al televisor, producción, ventas).

Variables categóricas

- Tienen características no medibles (nombres, etiquetas, colores).
- Asigna etiquetas o nombra las categorías de los objetos.

Escalas de Medición

- La escala de medición es una caracterización univoca de los objetos y generan un dato.
- Las escalas son nominales, ordinales, de intervalo y de razón.

Escalas Nominales

- Las escalas nominales carecen de orden, usan variables categóricas donde las etiquetas carecen de un sentido de magnitud solo identifican categorías.
- Pueden ser tanto numéricas como categóricas.

- Las escalas nominales **numéricas** son: los códigos postales; números telefónicos, numero asignado a las camisetas deportivas, colores a categorías de productos.
- Las escalas nominales **categóricas** son: tipo de sangre, grupo étnico, religión.

Escala Ordinal

- En una escala ordinal los valores registrados solo tienen un sentido de orden, la distancia entre un valor y otro no es constante, no hay un cero absoluto de referencia.

- Por ejemplo, las calificaciones asignadas a los estudiantes por letras: A, B, C, D y E. Cada una de estas letras indica un nivel de desempeño.
- Los grados académicos: licenciatura, maestría, doctorado.
- La calificación de la satisfacción de un producto o servicio otorgada por un cliente en una escala de 1 a 5.
- No es posible determinar la diferencia entre cada categoría de la escala.

Escala de Intervalo

- En una escala de intervalos los valores registrados tienen un sentido de orden, la distancia entre dos valores es constante, pero no hay un cero absoluto de referencia.

- Puntajes de la prueba de inteligencia
- Temperatura
- Índice de Precios del consumidor
- Índice de Pobreza

Escala de Razón

- En la escala de razón los valores registrados tienen un sentido de orden, la distancia entre dos valores es constante y además hay un cero absoluto de referencia.

- Peso en kilogramos de una persona
- Estatura en centímetros de un niño
- Salario mensual de un trabajador
- Edad en años de un empleado de Walmart
- Numero de paquetes enviados por una empresa

Escalas y Variables de Medición

Escala	Variable Categórica	Variable Cuantitativa
Nominal	SI	SI
Ordinal	SI	SI
Intervalo	NO	SI
Razón	NO	SI

Histogramas

- Cuando se tiene bastante información es muy abrumador *visualizar* los datos, por ello es necesario presentarlos en forma gráfica.
- La representación gráfica ayuda a observar la distribución de los datos de manera agrupada.
- Las gráficas son útiles para generar informes para que la información pueda ser aprovechada.
- Veremos la representación gráfica para variables **cuantitativas** y **categóricas**

Construcción de una Tabla de Distribución de Frecuencias

- **Primer paso:** se hace un arreglo de datos, esto es ponemos en orden de magnitud ascendente o descendente

Número de eventos de violencia doméstica reportados en un municipio				
10	14	21	22	17
15	14	18	33	23
20	15	19	16	28
22	27	18	18	13

Construcción de una Tabla de Distribución de Frecuencias

- **Segundo paso:** se calcula el rango de los datos. El rango es la distancia máxima entre el valor grande y el chico

Número de eventos de violencia doméstica reportados en un municipio				
10	15	18	20	23
13	15	18	21	27
14	16	18	22	28
14	17	19	22	33

$$\text{Rango} = R = \text{máx} - \text{mín} = 33 - 10 = 23$$

Construcción de una Tabla de Distribución de Frecuencias

- **Tercer paso:** Se calcula el número de intervalos de las clase necesarias. Usualmente se seleccionan entre 5 y 20.

$$K = 1 + 3.322 \log(n)$$

- donde n es el número de datos. Es decir,

$$K = 1 + 3.322 \log(20) = 1 + 4.32 = 5.32$$

- Utilizamos la función piso para obtener el valor, donde

$$\lceil K \rceil = \text{mín}\{n \in \mathbb{Z} \mid n \geq K\}$$

- Es decir, redondeamos hacia abajo, por lo que tenemos 5 intervalos.
- Otra forma de determinar el número de intervalos es obteniendo la raíz cuadrada del número de observaciones.

$$K = \sqrt{n} = \sqrt{20} = 4.47 \approx 5$$

Construcción de una Tabla de Distribución de Frecuencias

- **Cuarto paso:** determinar el ancho del intervalo

$$W = \frac{\text{Rango}}{\text{Clases}} = \frac{23}{5} = 4.6 \approx 5$$

- **Quinto paso:** es determinar las clases en si. Es decir los límite superior e inferior de cada intervalo.

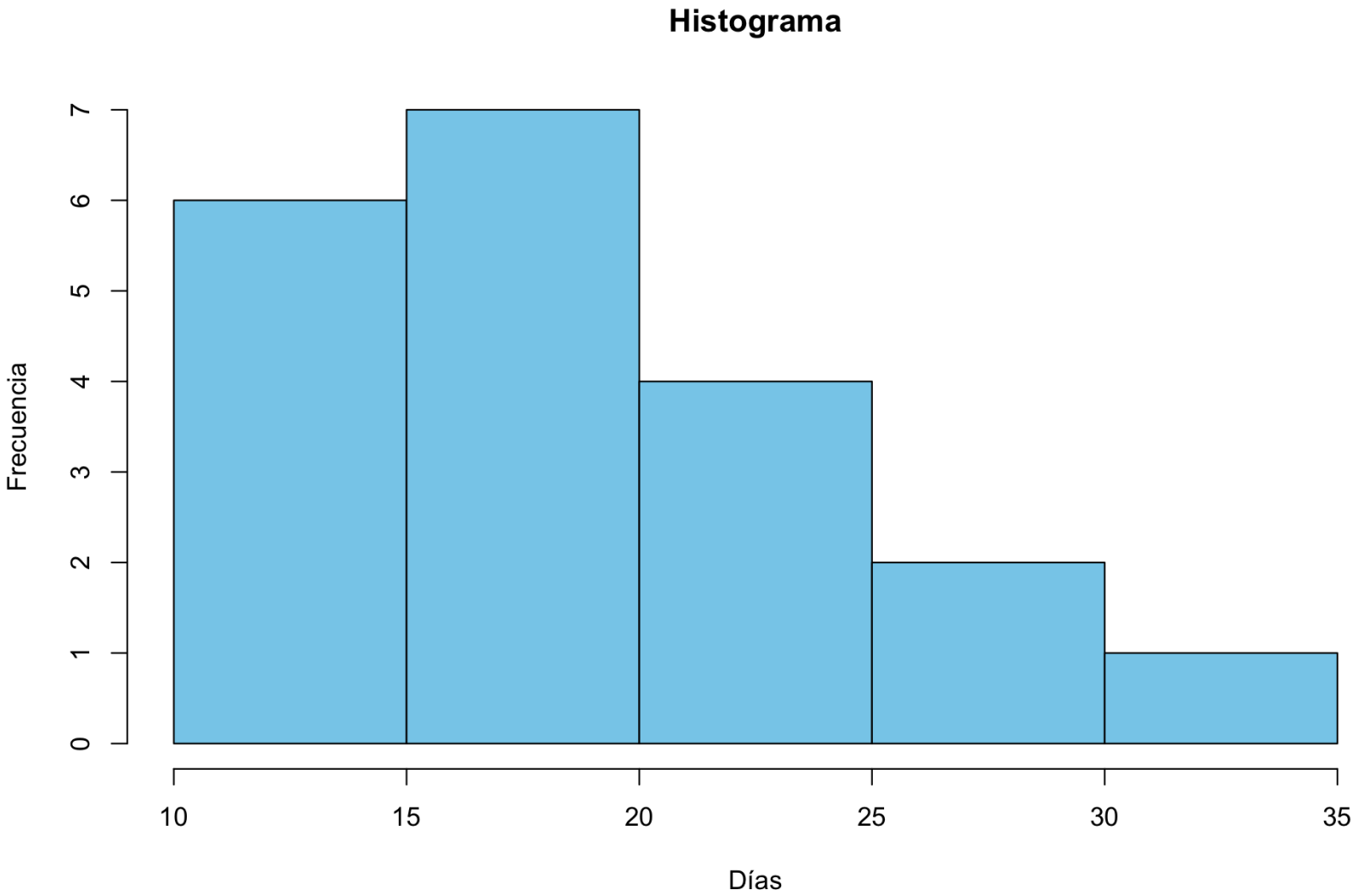
Clases	
10	15
15	20
20	25
25	30
30	35

Construcción de una Tabla de Distribución de Frecuencias

- El porcentaje acumulado se obtiene dividiendo las frecuencias acumuladas entre el total. Alternativamente, se pueden sumar los porcentajes relativos.

Clases	Frecuencias	Frecuencias Acumuladas	Porcentaje Relativo	Porcentaje Acumulado
(10, 15]	6	6	0.30	0.30
(15, 20]	7	13	0.35	0.65
(20, 25]	4	17	0.20	0.85
(25, 30]	2	19	0.10	0.95
(30, 35]	1	20	0.05	1.00
Total	20		1.00	

Histograma y Medidas Descriptivas



Medidas Descriptivas

Días	Descriptivas	Estadístico	Error Estándar
	Media	19.15	1.2508
	Mediana	18	
	Varianza	31.2921	
	Máx.	33	
	Mín.	10	
	Rango	23	
	Skewness	0.7093	0.6925
	Kurtosis	-0.0559	-0.0281

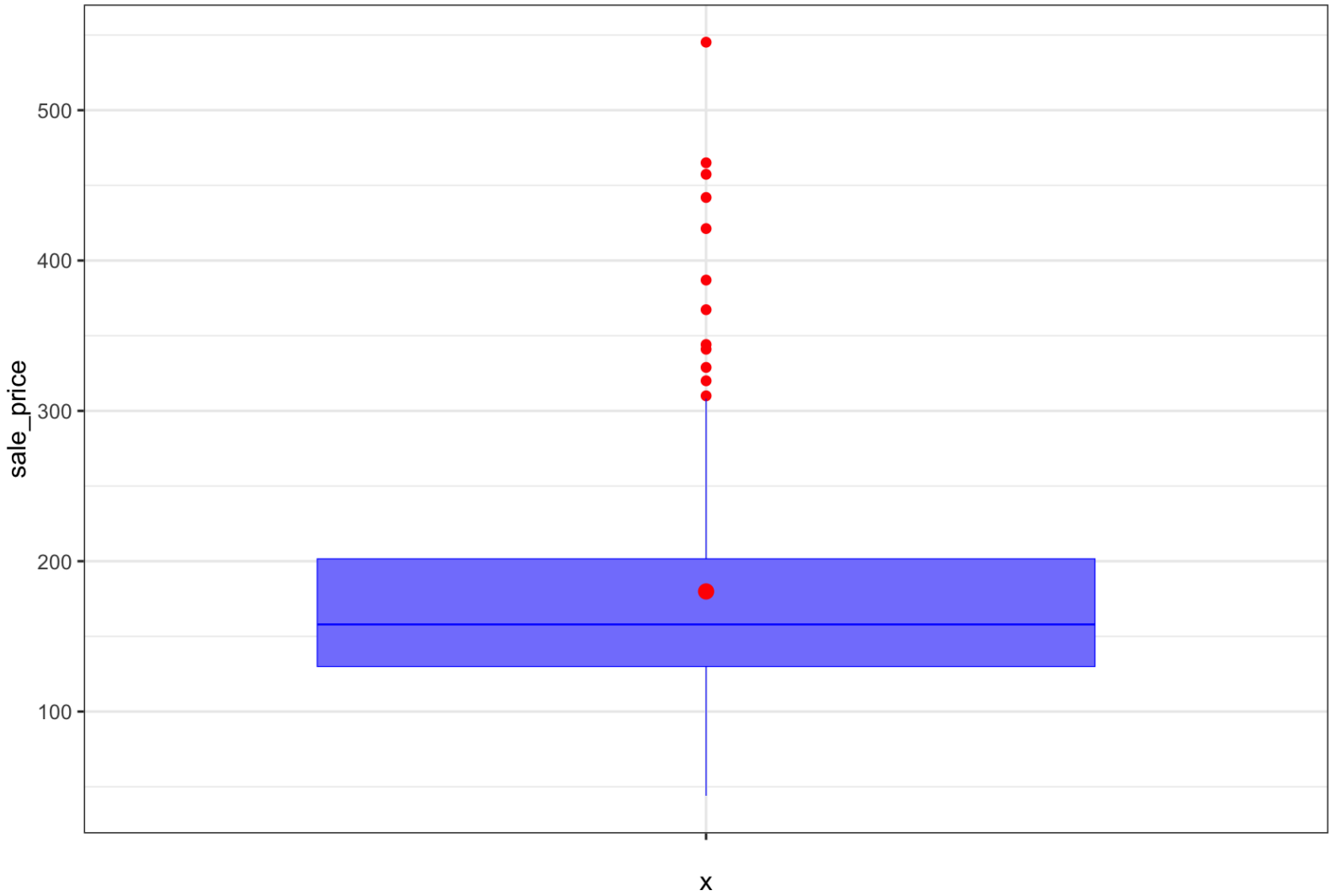
Código en R

```
library(ggplot2)
library(pastecs)

data <- c(10, 14, 21, 22, 17, 15, 14, 18, 33, 23,
          20, 15, 19, 16, 28, 22, 27, 18, 18, 13)

hist(data, breaks = seq(min(data), max(data) + 5, by = 5),
       main = "Histograma",
       xlab = "Dias",
       ylab = "Frecuencia",
       col = "skyblue",
       border = "black")

summary(data)
stat.desc(data, norm=TRUE)
```

El Diagrama de Caja

- **Límite superior.** Es el extremo superior de la caja. Los puntajes por encima de este límite se consideran atípicas.
- **Tercer cuartil (Q_3):** Por debajo de este valor se encentran como máximo el 75 % de las opiniones de los estudiantes.
- **Mediana:** Coincide con el segundo cuartil. Divide a la distribución en dos partes iguales. De este modo, 50 % de las observaciones están por debajo de la mediana y 50 % está por encima.
- **Primer cuartil (Q_1):** Por debajo de este valor se encuentra como máximo el 25 % de las opiniones de los estudiantes
- **Límite inferior:** Es el extremo inferior del bigote. Las opiniones por debajo de este valor se consideran atípicas.
- **Valores atípicos:** Opiniones que están apartadas del cuerpo principal. Pueden ser efectos de causas extrañas, opiniones extremas o en el caso de la tabulación manual, errores de medición o registro.

Diagrama de Caja en R



Código en R

```
library(tidyverse)
library(r02pro)

sale_price <- na.omit(sahp$sale_price)
summary(sale_price)

ggplot(data = sahp, aes(x = "", y = sale_price)) +
  geom_boxplot(fill = "#8484fc", color = "blue", outlier.
 color = "red", lwd=0.2) +
  geom_point(stat = "summary",
             fun = "mean",
             shape = 20,
             size = 4,
             color = "red")
```

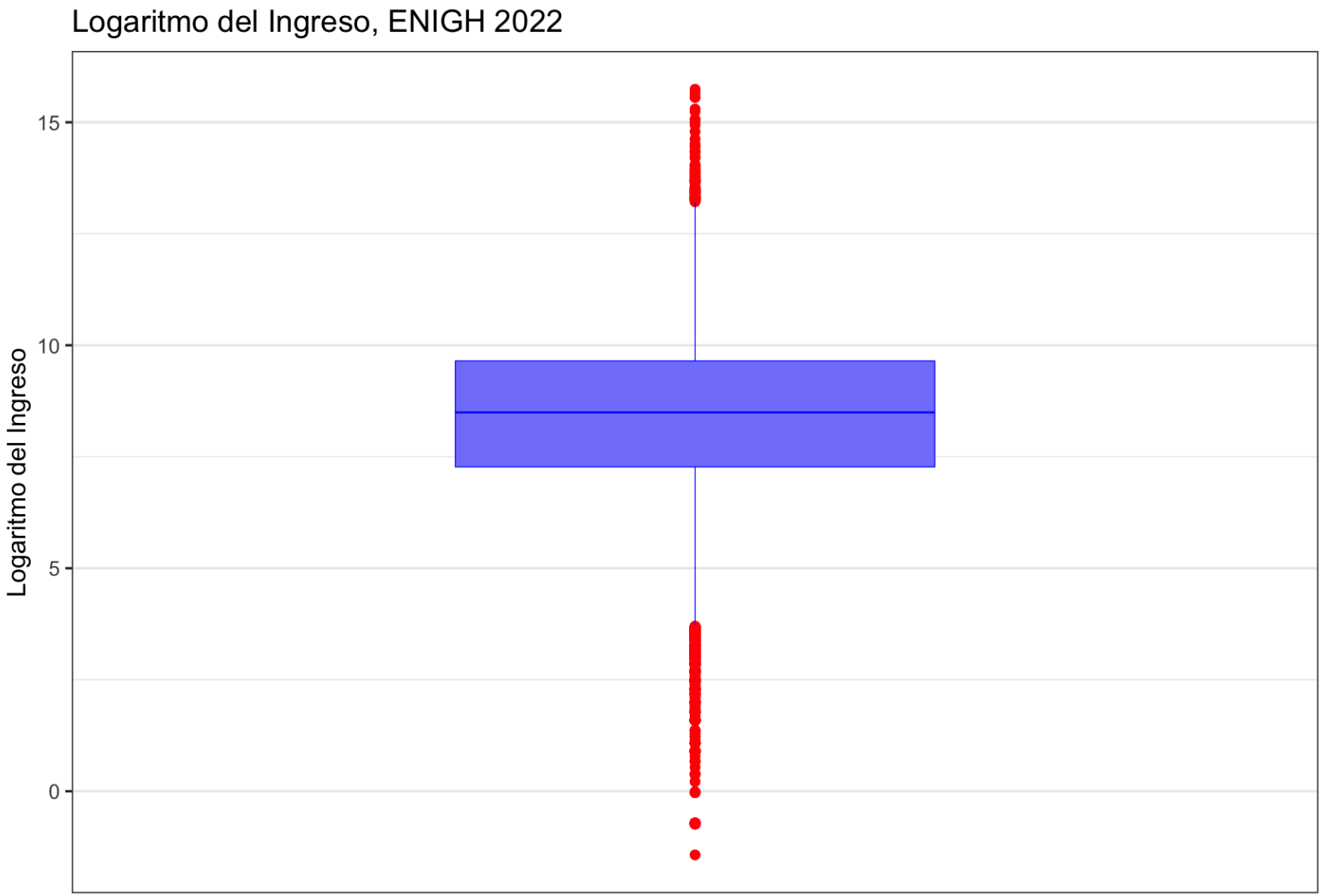
Algunas consideraciones

- La mediana puede inclusive coincidir con los cuartiles o con los límites de los bigotes. Esto sucede cuando se concentran muchos datos en un mismo punto. Pudiera ser este un caso particular de una distribución sesgada o el caso de una distribución muy homogénea.
- En este caso, las medidas descriptivas que representa el gráfico de caja, y que podemos ver en la figura anterior, son las siguientes:

Primer cuartil	Tercer cuartil	Mediana	Media
130.0	201.6	157.9	179.9

- A continuación, veremos otro ejemplo.

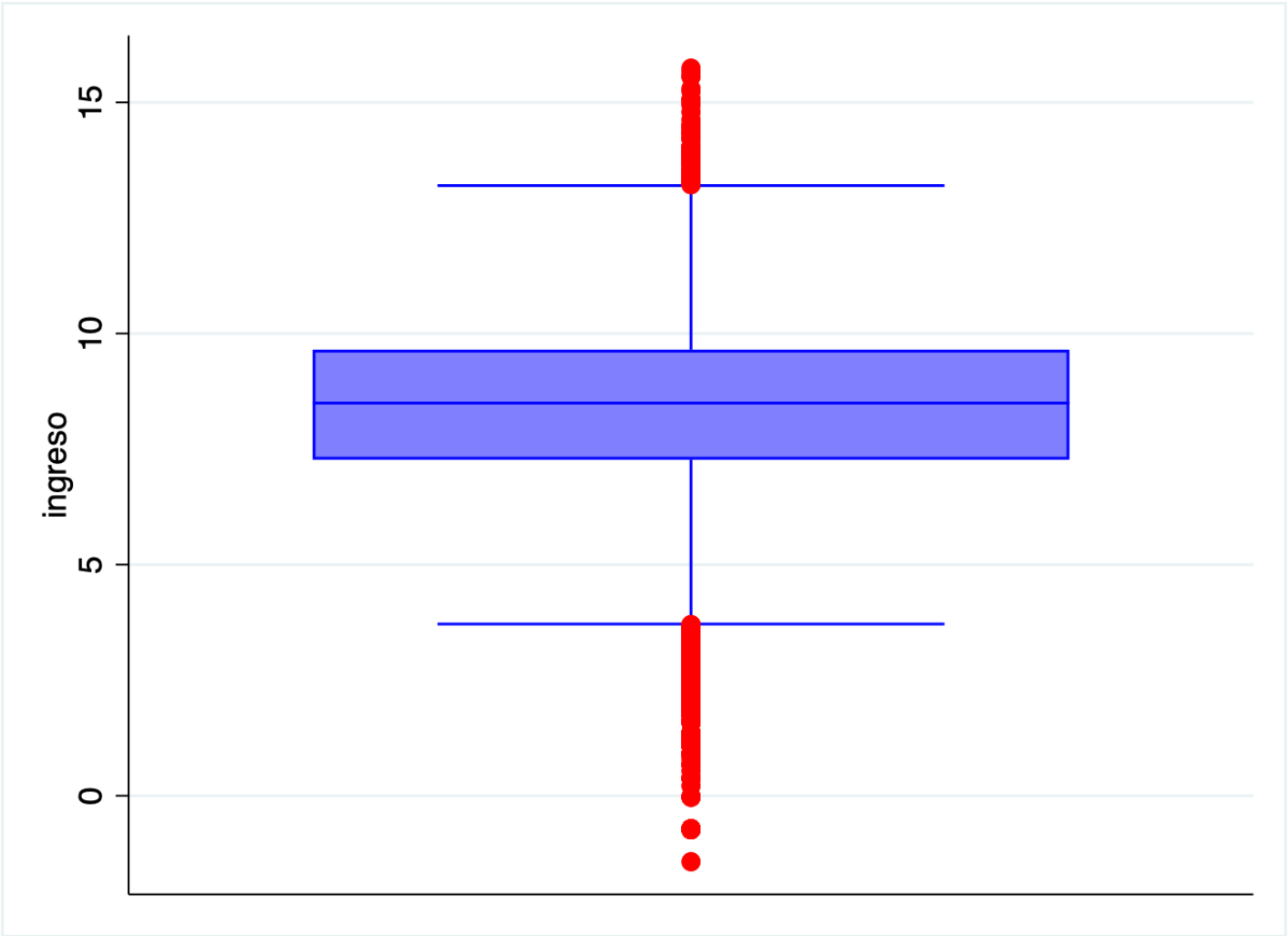
Diagrama de Caja en R



Código en R

```
library(ggplot2)
theme_set(theme_bw())

log_ingreso <- log(database$ing_tri)

ggplot() +
  geom_boxplot(aes(y = log_ingreso), fill = "#8484fc",
 color = "blue", outlier.color = "red") +
  scale_x_discrete() +
  labs(title = "Logaritmo del Ingreso, ENIGH 2022",
       y = "Logaritmo del Ingreso")
```

Diagrama de Caja en Stata



Bibliografía

Bhattacharyya, G. K., y Johnson, R. A. (1977). *Statistical concepts and methods*. Wiley.

Bickel, P. J., y Doksum, K. A. (2015). *Mathematical statistics: basic ideas and selected topics, volumes i-ii package*. CRC Press.

Canavos, G. C., y Medal, E. G. U. (1987). *Probabilidad y estadística*. McGraw Hill México.

Mendenhall, W. (2002). *Introducción a la probabilidad y estadística; edit.* Thomson México;.

Wackerly, D., Mendenhall, W., y Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

Wackerly, D. D., Mendenhall III, W., Scheaffer, R. L., y Milanés, Y. (2002). *Estadística matemática con aplicaciones*. Cengage Learning Editores.

¡Gracias por su atención!